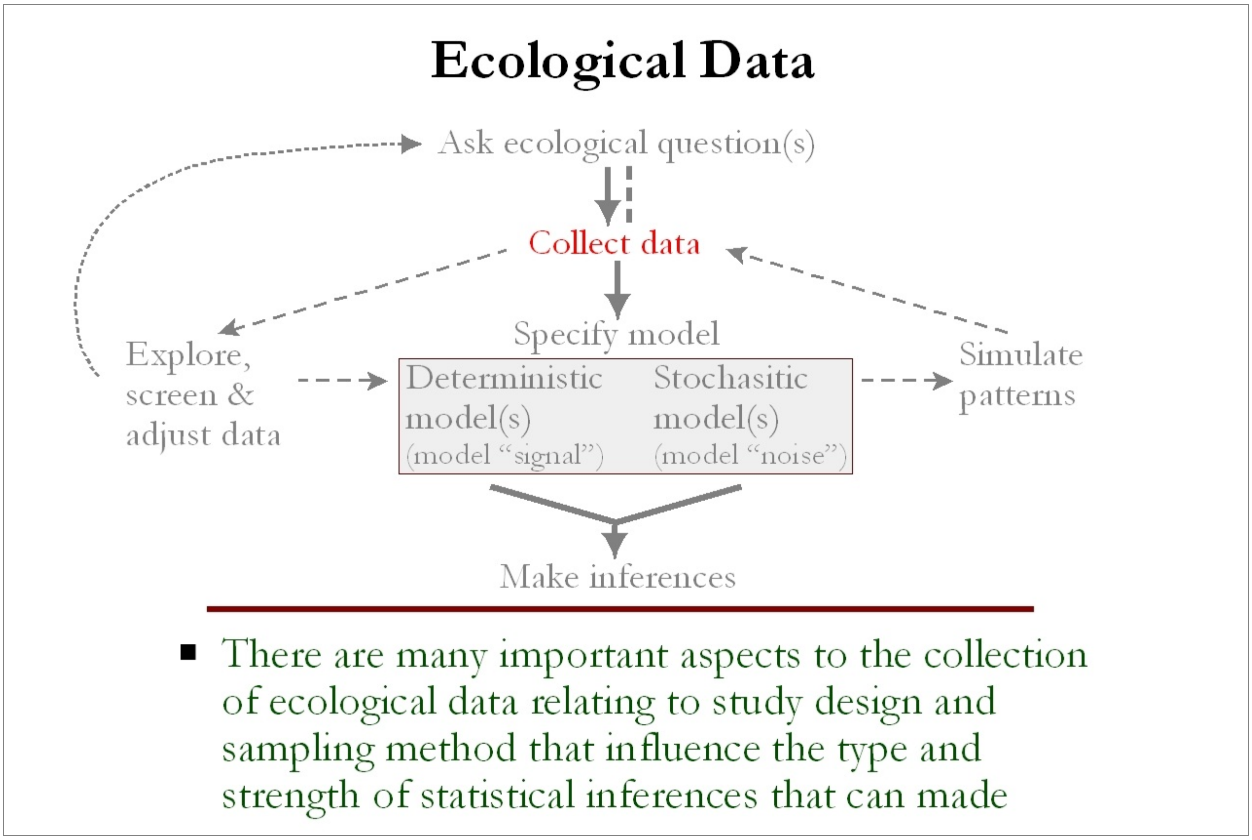
Design and Analysis of Ecological Data
Conceptual Foundations:
Ecological Data



Ecological Data

Ask ecological question(s)

Collect data

Explore, screen & adjust data

Specify model

Deterministic model(s) (model "signal")

Stochasitic model(s) (model "noise")

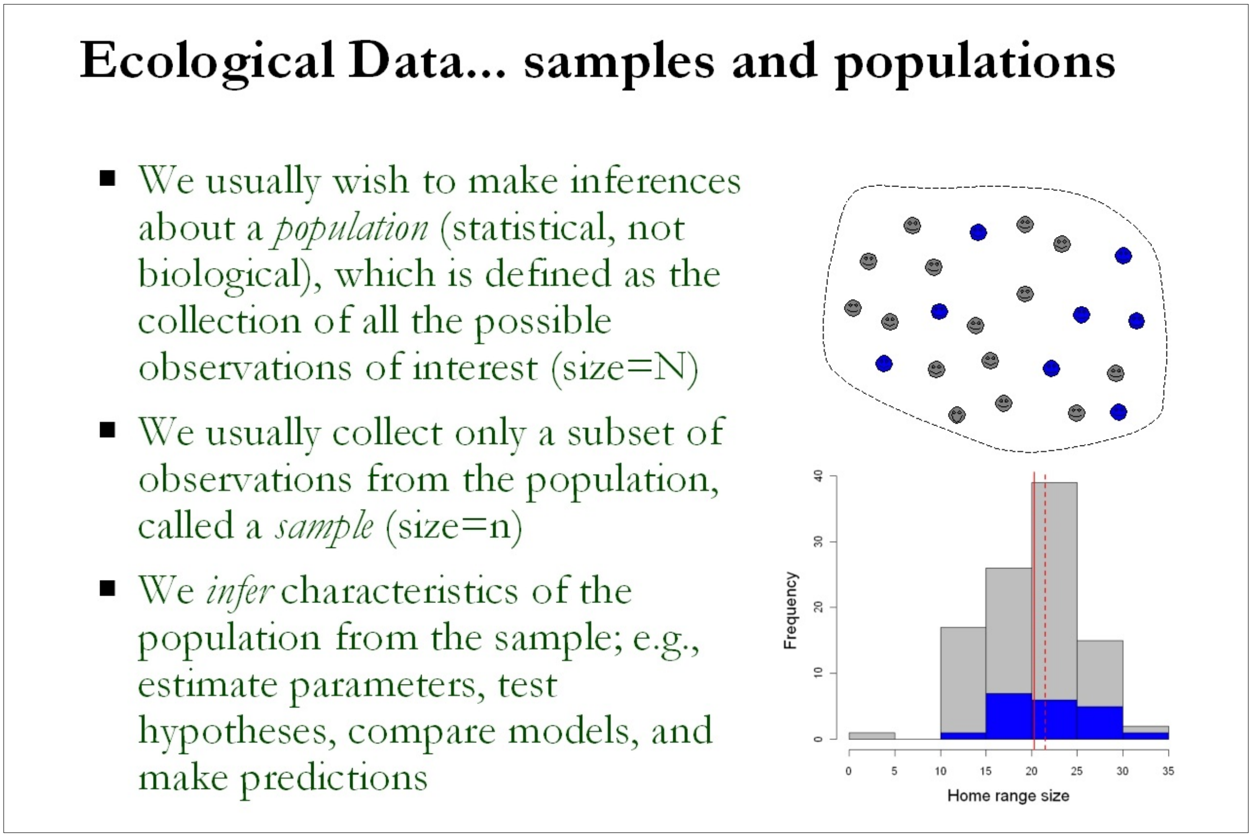
Simulate patterns

Make inferences

- There are many important aspects to the collection of ecological data relating to study design and sampling method that influence the type and strength of statistical inferences that can made

1. Purpose of data collection

Ideally, once the ecological question has been identified, the study is designed and the data is collected in a manner that will result in strong inferences. There are many important aspects to the collection of ecological data relating to study design and sampling method that will influence the type and strength of statistical inferences that can made: identifying the desired scope of inference, choosing appropriate observational/experimental units, choosing the types of data to collect, and establishing a robust sampling scheme (i.e., spatial and/or temporal distribution of units and method(s) of collecting the data) to ensure accurate and precise inferences. We will consider some of these issues in the second part of this course. Here we will briefly distinguish samples from populations, describe the major different types of ecological data and some of the properties of each type, the types of variables and their relationships, and preview some of the important issues to consider in study design that we will discuss further in the second part of the course.



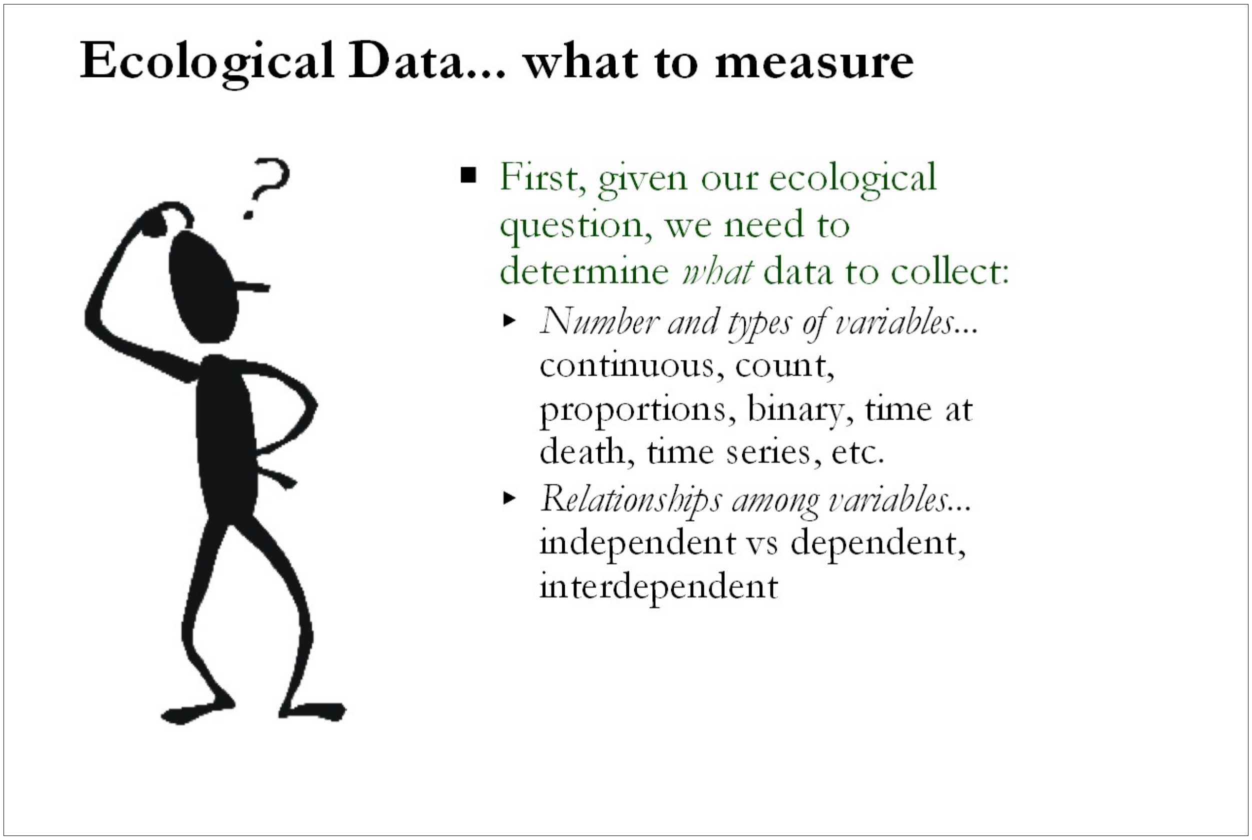
Ecological Data... samples and populations

- We usually wish to make inferences about a *population* (statistical, not biological), which is defined as the collection of all the possible observations of interest (size=N)
- We usually collect only a subset of observations from the population, called a *sample* (size=n)
- We *infer* characteristics of the population from the sample; e.g., estimate parameters, test hypotheses, compare models, and make predictions

2. Samples and populations

We usually wish to make inferences about a population (statistical, not biological), which is defined as the collection of all the possible observations of interest. A biological population under consideration may or may not constitute the statistical population if, for example, the functional population extends over a broader geographic extent than the study area. We usually represent the size of the statistical population in statistical formulae as upper case N. For lots of practical reasons, we usually collect only a subset of observations from the population, and we represent the size of the sample as lower case n.

Importantly, we *infer* characteristics of the population from the sample; e.g., estimate parameters, test hypotheses, compare models, and make predictions. Thus, the entire realm of inferential statistics applies when we seek to draw conclusions from a sample about the underlying population. Otherwise, we may be interested in or forced to merely describe the patterns in the sample without explicit inference to the population – the realm of descriptive statistics. Note, in rare cases, we may actually observe every possible entity of interest – the population – in which case simple descriptive statistics suffice to draw conclusions from since we know with exactness (to the precision of our measurement system) the characteristics of the population we are studying.



Ecological Data... what to measure

- First, given our ecological question, we need to determine *what* data to collect:
  - *Number and types of variables...* continuous, count, proportions, binary, time at death, time series, etc.
  - *Relationships among variables...* independent vs dependent, interdependent

3. Types of data

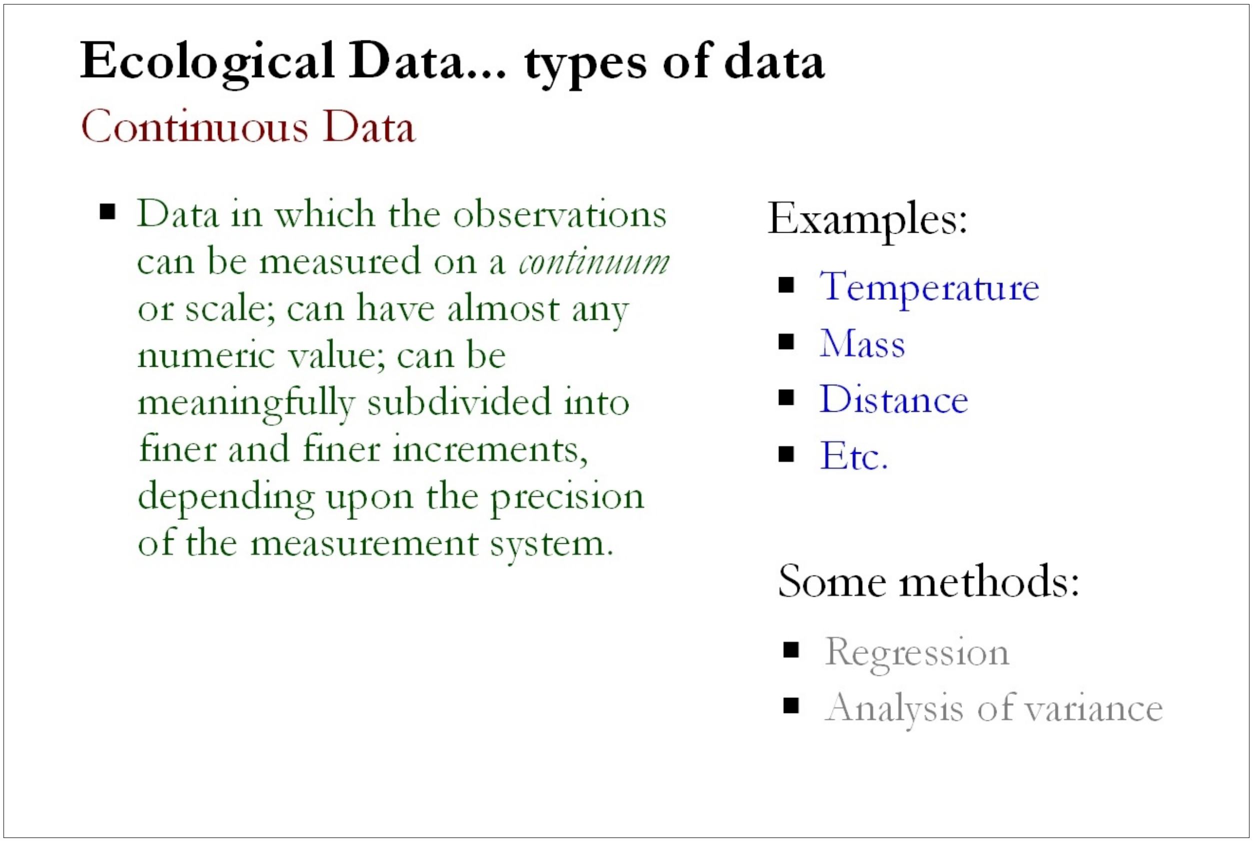
Once we have identified our ecological question, the first thing we need to do is determine what data to collect. This is one of the most important steps in the entire modeling process, because if we collect the wrong type of data, no statistical model of any kind will allow us to answer our ecological question. While there are many important considerations to this step, we need to carefully consider the number and types of variables to collect and their relationships.

In ecological studies, there are several major types of data:

- continuous data
- counts
- proportions
- binary data
- time at death
- time series
- circular data

And there are at least three major types of variables based on their relationships to each other:

- independent variables
- dependent variables
- interdependent variables



Ecological Data... types of data

Continuous Data

- Data in which the observations can be measured on a *continuum* or scale; can have almost any numeric value; can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system.

Examples:

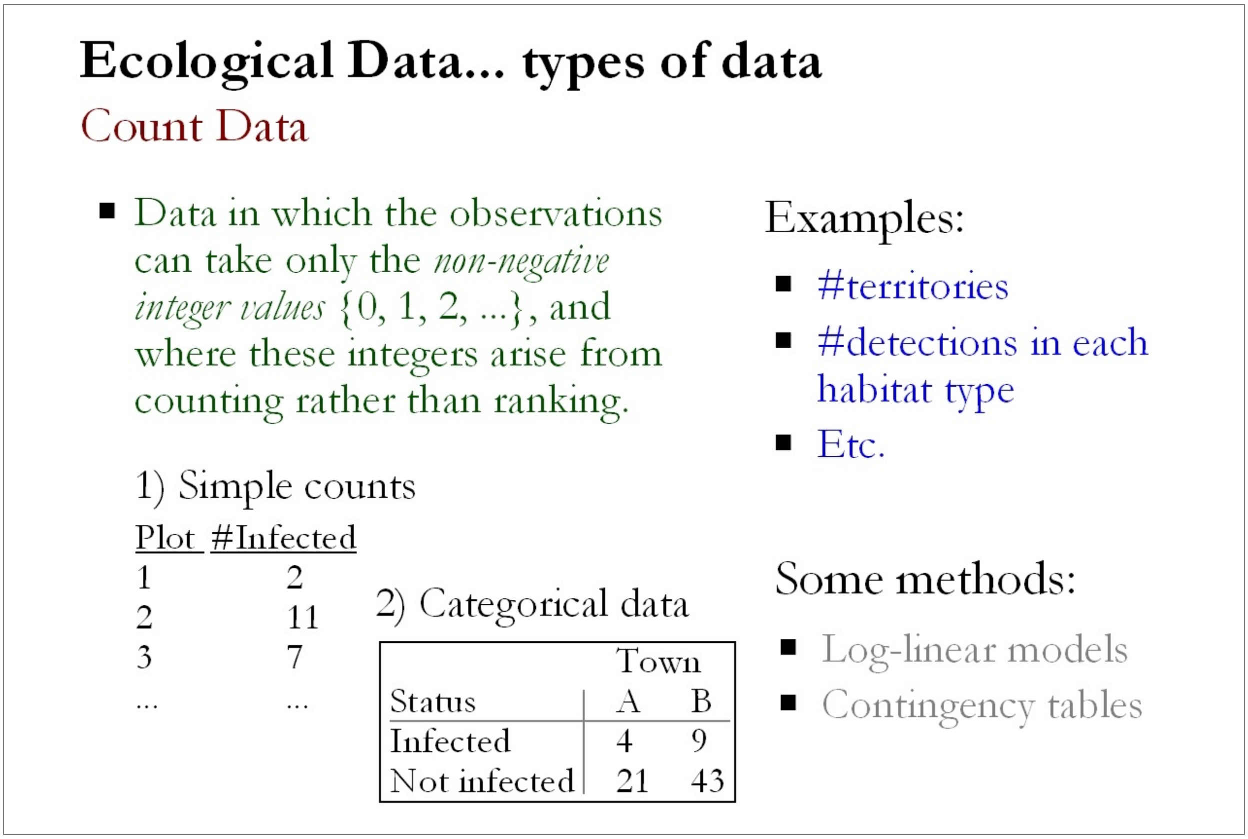
- Temperature
- Mass
- Distance
- Etc.

Some methods:

- Regression
- Analysis of variance

3.1 Continuous data

Continuous data is data in which the observations can be measured on a *continuum* or scale; can have almost any numeric value; and can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system. There are lots of examples of continuous data: temperature, mass, distance, etc. This is the most common type of ecological data collected and there are lots of statistical methods designed to work with this type of data, such as regression and analysis of variance.



Ecological Data... types of data

Count Data

- Data in which the observations can take only the *non-negative integer values* {0, 1, 2, ...}, and where these integers arise from counting rather than ranking.

1) Simple counts

Plot	#Infected
1	2
2	11
3	7
...	...

2) Categorical data

	Town	
Status	A	B
Infected	4	9
Not infected	21	43

Examples:

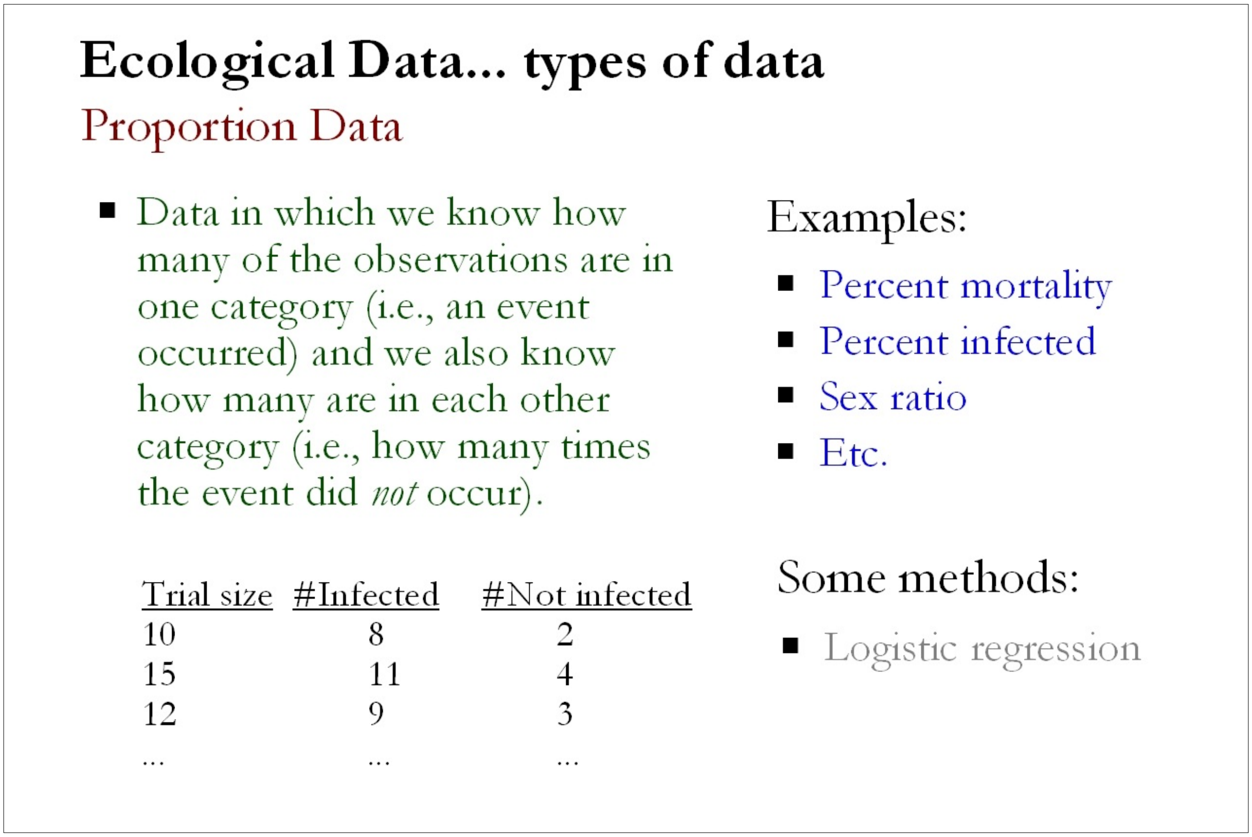
- #territories
- #detections in each habitat type
- Etc.

Some methods:

- Log-linear models
- Contingency tables

3.2 Count data

Count data is a form of discrete data in which the observations can take only the *non-negative integer values* {0, 1, 2, ...}, and where these integers arise from counting rather than ranking. Count data is usually of one of two forms: 1) simple counts, e.g., the number of plants infected by a disease on a plot, the number of eggs in a nest, etc., and 2) categorical data, in which the counts are tallied for one or more categorical explanatory variables, e.g., the number of plants infected in each of several towns. With simple counts, the goal is usually to explain or predict the counts based on one or continuous independent or explanatory variables, and the method of generalized linear modeling is used for this purpose. With categorical data, the goal is usually to determine whether the distribution of counts among categories differs from expected, and the method of contingency table analysis employing log-linear modeling is often used for this purpose.



Ecological Data... types of data

Proportion Data

- Data in which we know how many of the observations are in one category (i.e., an event occurred) and we also know how many are in each other category (i.e., how many times the event did *not* occur).

Trial size	#Infected	#Not infected
10	8	2
15	11	4
12	9	3
...	...	...

Examples:

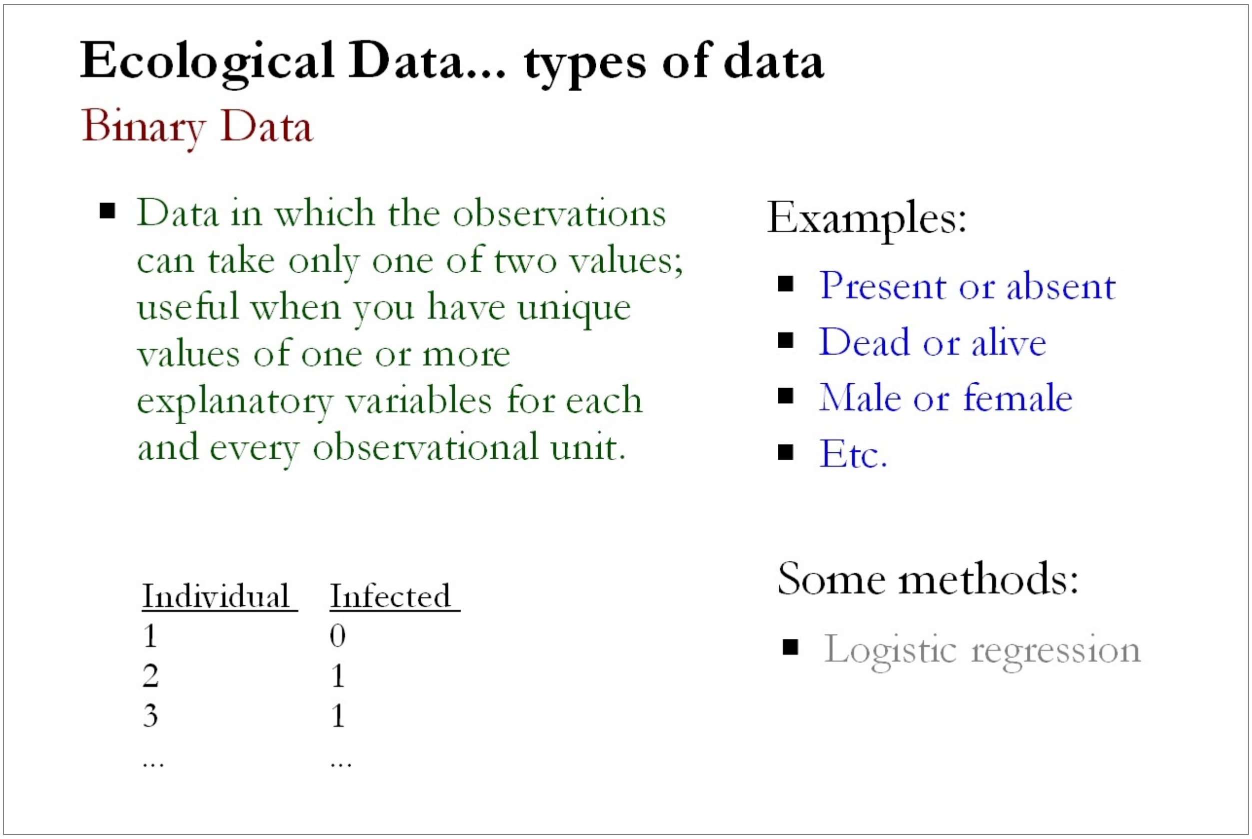
- Percent mortality
- Percent infected
- Sex ratio
- Etc.

Some methods:

- Logistic regression

3.3 Proportion data

Proportion data is another form of discrete data in which we know how many of the observations are in one category (i.e., an event occurred) and we also know how many are in each other category (i.e., how many times the event did *not* occur). This is an important distinction, since it allows the data to be represented as proportions instead of frequencies, as with count data. There are lots of ecological examples of proportion data: percent mortality, percent infected, sex ratio, etc.. The key distinction of proportion data is that the frequency of the event, e.g., individual died, is known as well as the total number of events, e.g., total number of individuals. With proportion data, the goal is typically to explain or predict the proportional response based on one or more explanatory variable, and the method of logistic regression is designed for this purpose. Note, here the explanatory variables are measured for each sample trial, as opposed to each individual. This is an important distinction between proportion data and binary data (next).



Ecological Data… types of data

Binary Data

- Data in which the observations can take only one of two values; useful when you have unique values of one or more explanatory variables for each and every observational unit.

Individual	Infected
1	0
2	1
3	1
…	…

Examples:

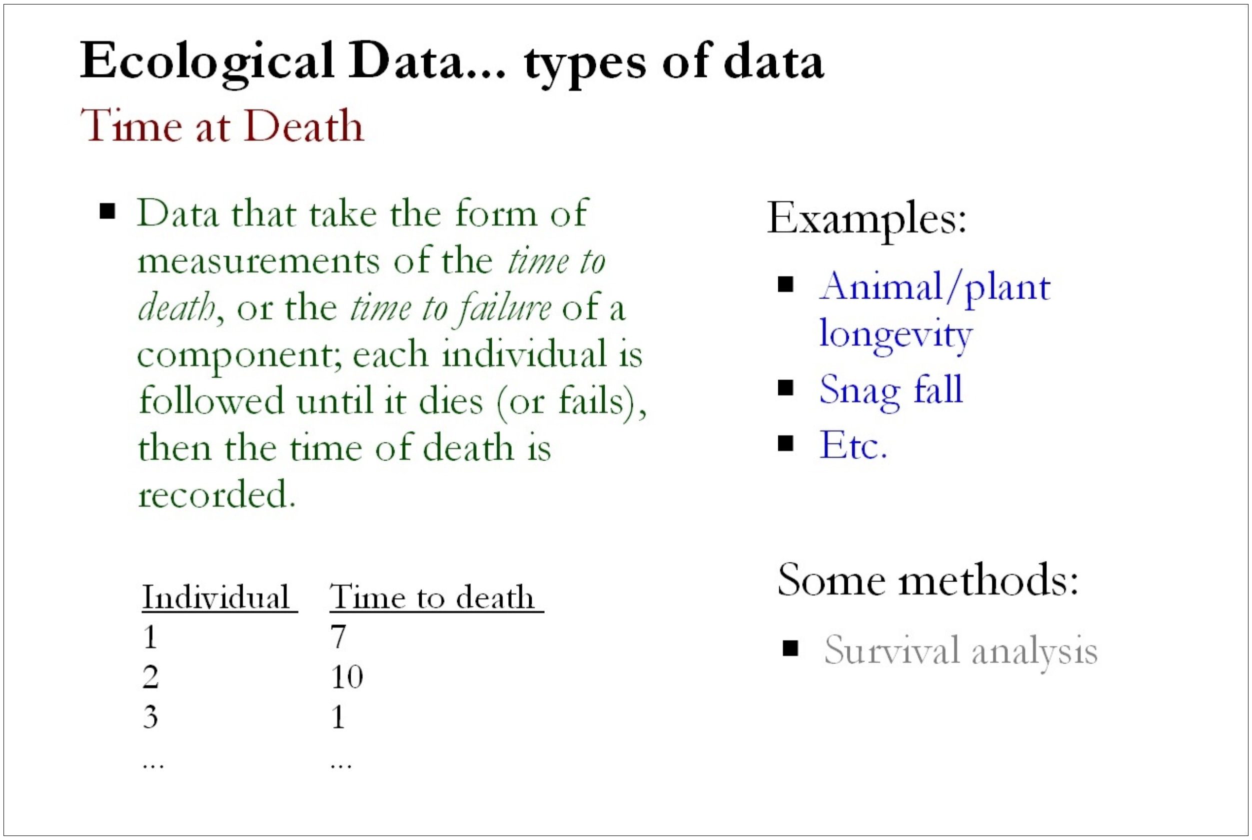
- Present or absent
- Dead or alive
- Male or female
- Etc.

Some methods:

- Logistic regression

3.4 Binary data

Binary data is data in which the observations can take only one of two values, for example, alive or dead, present or absent, male or female, etc.. Binary data is useful when you have unique values of one or more explanatory variables for each and every observational unit; this is an important distinction from proportional data in which the explanatory data is collected at the level of the trial (consisting of many observational units). Binary data is typically analyzed with the method of logistic regression, like proportion data.



Ecological Data... types of data

Time at Death

- Data that take the form of measurements of the *time to death*, or the *time to failure* of a component; each individual is followed until it dies (or fails), then the time of death is recorded.

Individual	Time to death
1	7
2	10
3	1
...	...

Examples:

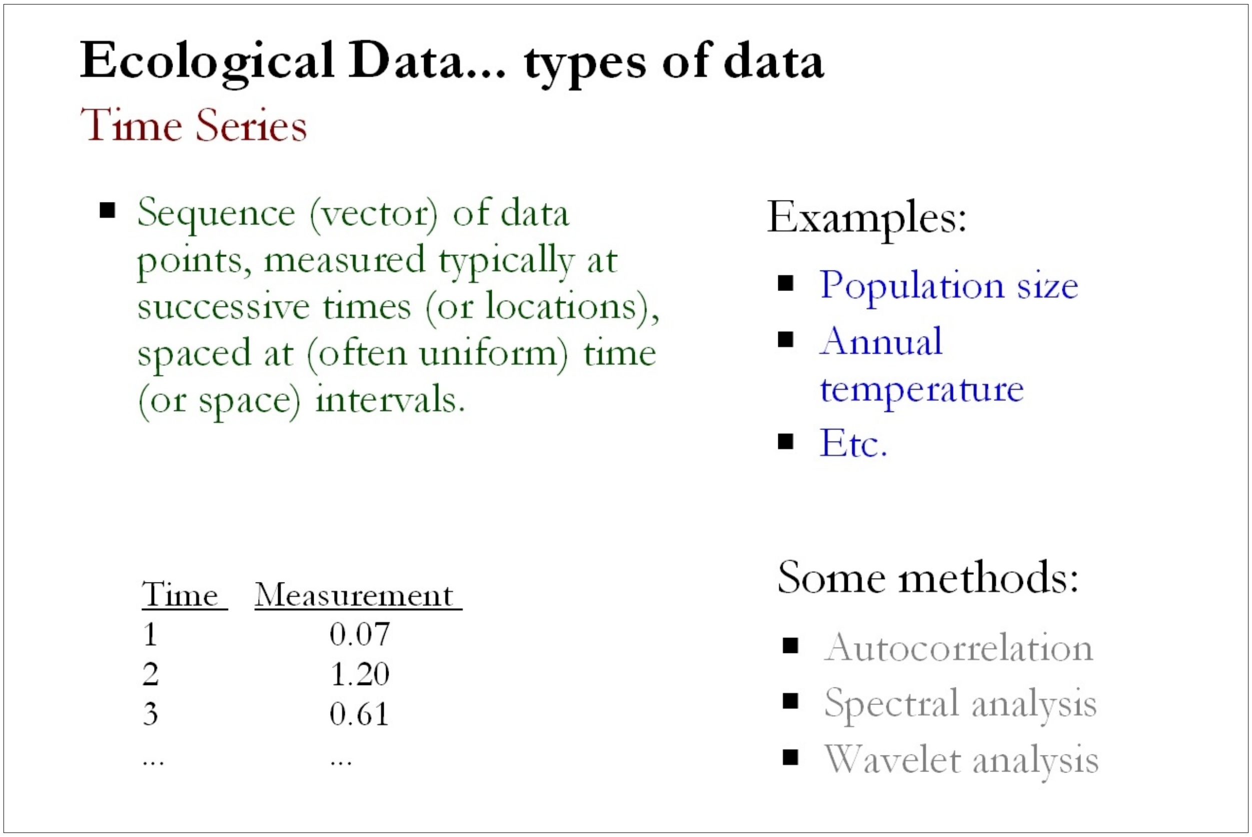
- Animal/plant longevity
- Snag fall
- Etc.

Some methods:

- Survival analysis

3.5 Time at death data

Time at death data is data that take the form of measurements of the *time to death*, or the *time to failure* of a component; each individual is followed until it dies (or fails), then the time of death (or failure) is recorded. Time at death data is not limited to plant and animal longevity studies, however, it applies to any situation in which the time to completion of a process is relevant.; for example, the time it takes juveniles to disperse out of the study area, or the time it takes a snag to fall. Time at death data is analyzed by the special method of survival analysis, which is highly complex and rapidly evolving to account for all sorts of variations.



Ecological Data... types of data

Time Series

- Sequence (vector) of data points, measured typically at successive times (or locations), spaced at (often uniform) time (or space) intervals.

Time	Measurement
1	0.07
2	1.20
3	0.61
...	...

Examples:

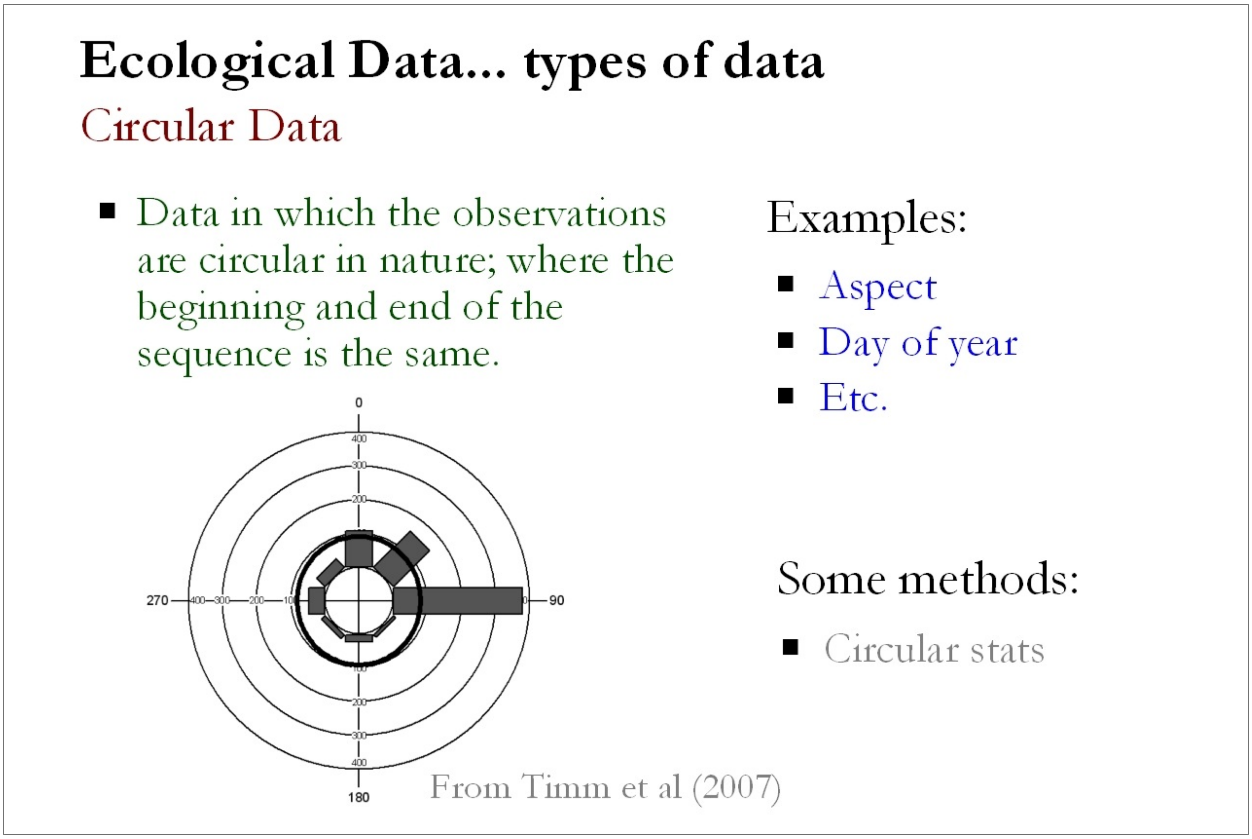
- Population size
- Annual temperature
- Etc.

Some methods:

- Autocorrelation
- Spectral analysis
- Wavelet analysis

3.6 Time series data

Time series data involves a sequence (vector) of data points, measured typically at successive times (or locations), spaced at (often uniform) time (or space) intervals. Usually time series data contains repeated patterns of variation, and identifying and quantifying the scale(s) of the repeated pattern is often the focus of the analysis. There are many examples of time series data in ecology: population size measured annually, temperature data measured at fixed intervals, river discharge measured over time, etc. And let's not forget that time series data also includes spatial data that is serially correlated in space rather than time, such as variables measured at intervals along transects, e.g., plant cover, soil chemistry, water depth, etc.. There several specialized analytical methods for time series data, include autocorrelation analysis, spectral analysis, and wavelet analysis to name just a few.



Ecological Data... types of data

Circular Data

- Data in which the observations are circular in nature; where the beginning and end of the sequence is the same.

Examples:

- Aspect
- Day of year
- Etc.

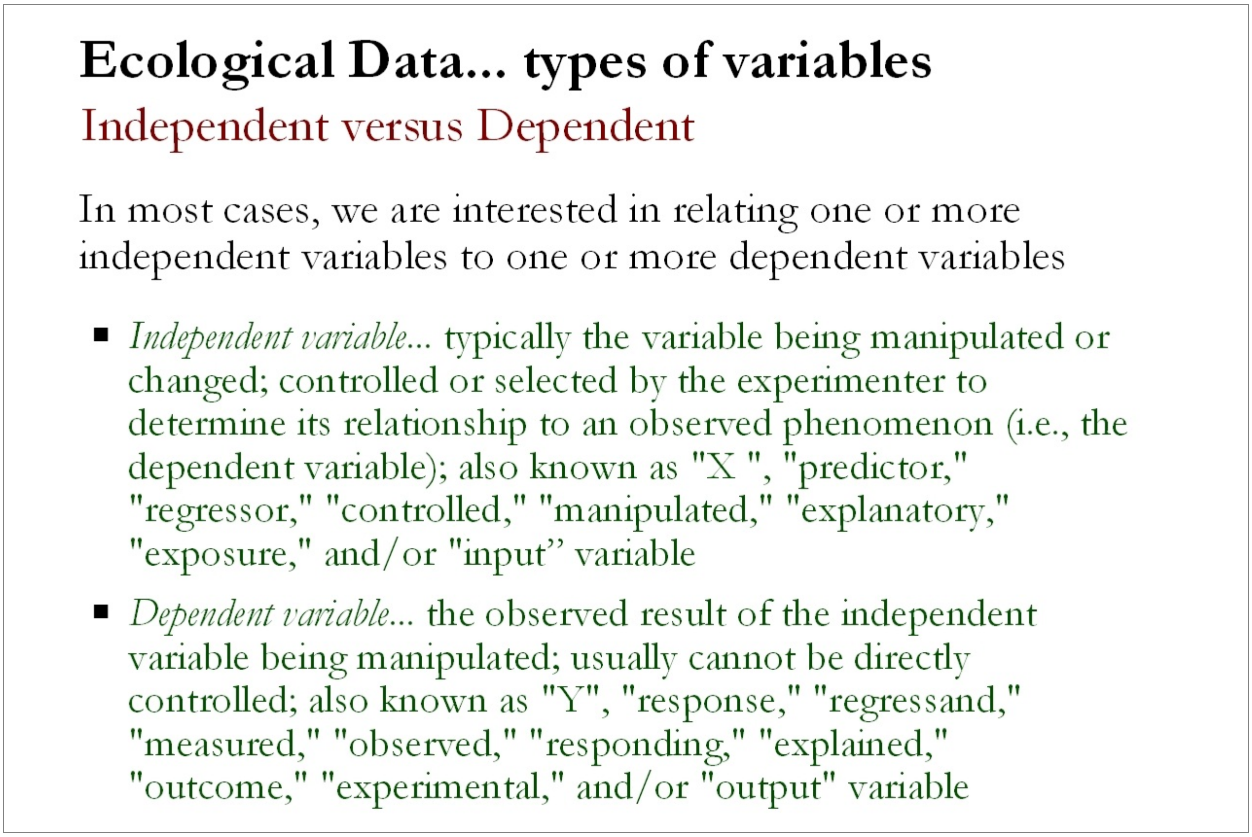
Some methods:

- Circular stats

From Timm et al (2007)

3.7 Circular data

Circular data is, not surprisingly, data in which the observations are circular in nature; where the beginning and end of the sequence is the same. Classic examples of circular data are topographic aspect, day of the year, and orientation of movement. The figure shown here is a circular histogram depicting red-spotted newt (Notophthalmus viridescens) departure from a uniform distribution for emigrating juveniles leaving a natal pond in western Massachusetts based on eight directional bins. In the histogram, each arm depicts one of the eight directional bins, concentric circles represent a given raw number of individuals, and the bold circle delineates the expected bin value given a uniform distribution. Circular data is typically analysed with specialized methods.



Ecological Data... types of variables

Independent versus Dependent

In most cases, we are interested in relating one or more independent variables to one or more dependent variables

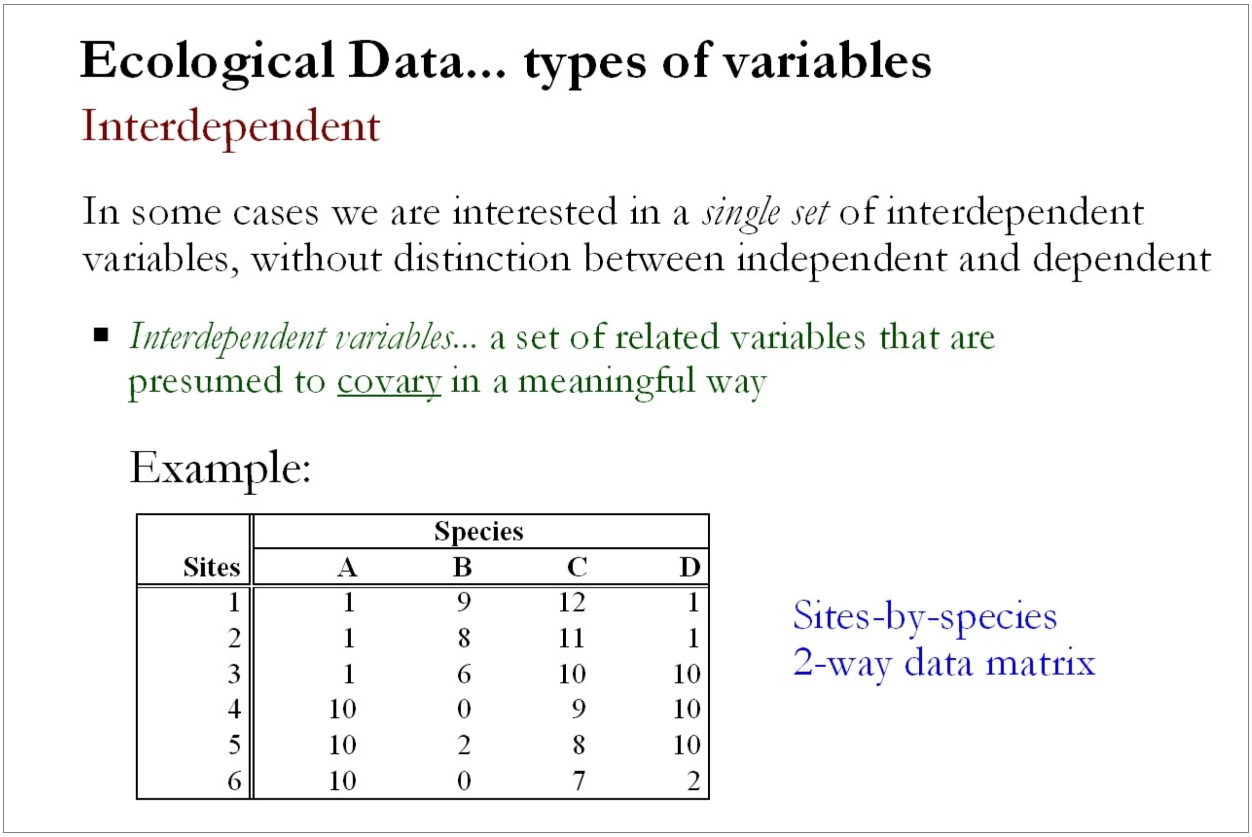
- *Independent variable...* typically the variable being manipulated or changed; controlled or selected by the experimenter to determine its relationship to an observed phenomenon (i.e., the dependent variable); also known as "X ", "predictor," "regressor," "controlled," "manipulated," "explanatory," "exposure," and/or "input" variable
- *Dependent variable...* the observed result of the independent variable being manipulated; usually cannot be directly controlled; also known as "Y", "response," "regressand," "measured," "observed," "responding," "explained," "outcome," "experimental," and/or "output" variable

4. Types of variables

In most, but not all, studies, our ecological question requires that we collect data on two or more variables in which one or more variables are considered as “independent” variables and one or more are considered as “dependent” variables. This distinction is critical to most statistical models.

Independent variable... typically the variable(s) being manipulated or changed, or the variable(s) controlled or selected by the experimenter to determine its relationship to an observed phenomenon (i.e., the dependent variable). In observational studies, the independent variable(s) is not explicitly manipulated or controlled through experimentation, but rather observed in its naturally occurring variation, yet it is presumed determine or influence the value of the dependent variable. The independent variable is also known as "x ", "explanatory," "predictor," "regressor," "controlled," "manipulated," "exposure," and/or "input" variable.

Dependent variable... the observed result of the independent variable(s) being manipulated, and it usually cannot be directly controlled. The dependent variable is generally the phenomenon whose behavior we are interested in understanding. The dependent variable is also known as "y", "response," "regressand," "measured," "observed," "responding," "explained," "outcome," "experimental," and/or "output" variable.



Ecological Data... types of variables

Interdependent

In some cases we are interested in a *single set* of interdependent variables, without distinction between independent and dependent

- *Interdependent variables...* a set of related variables that are presumed to covary in a meaningful way

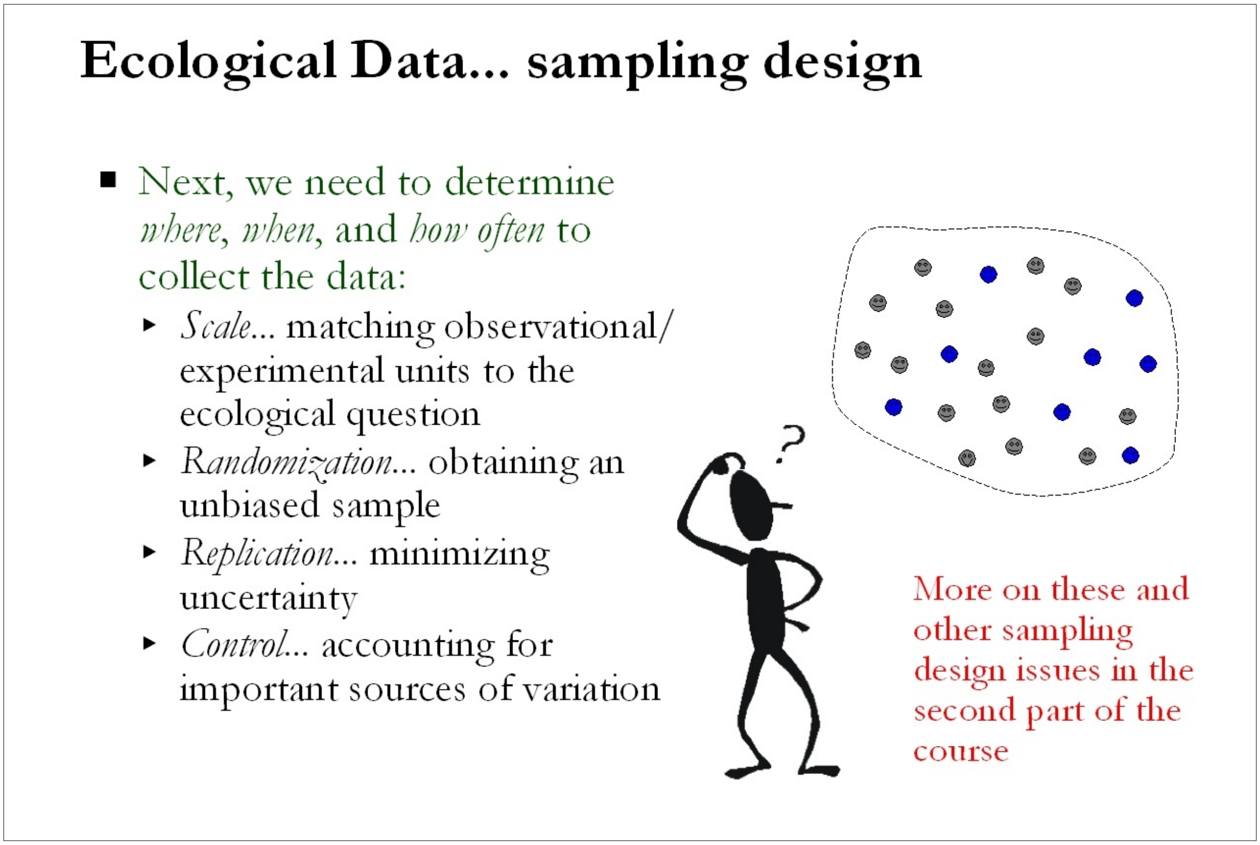
Example:

	Species			
Sites	**A**	**B**	**C**	**D**
1	1	9	12	1
2	1	8	11	1
3	1	6	10	10
4	10	0	9	10
5	10	2	8	10
6	10	0	7	2

Sites-by-species
2-way data matrix

In some cases we are interested in a *single set* of interdependent variables, without distinction between independent and dependent

Interdependent variables... a set of related variables that are presumed to covary in a meaningful way. A common example is a community data set consisting of *n* sites by *p* species abundances, arranged in a two-way data matrix in which the rows represent the sites and the columns represent the species. In this case, the species are the variables and there is no distinction of independent and dependent. In fact, they are all presumed to be interdependent on each other since they presumably covary in meaningful ways. Moreover, they are generally considered to be inter-*dependent* variables because they are presumed to respond to other perhaps unmeasured independent variables that are not part of this variable set.



Ecological Data... sampling design

- Next, we need to determine *where*, *when*, and *how often* to collect the data:
  - *Scale*... matching observational/ experimental units to the ecological question
  - *Randomization*... obtaining an unbiased sample
  - *Replication*... minimizing uncertainty
  - *Control*... accounting for important sources of variation

More on these and other sampling design issues in the second part of the course

5. Sampling Design

Once we have determined what data to collect to answer our ecological question, the next thing we need to is determine *where*, *when*, and *how often* to collect the data. This is complicated arena of sampling design and there are many critical issues to consider, such as:

- *Scale*... matching observational/ experimental units to the ecological question
- *Randomization*... obtaining an unbiased sample
- *Replication*... minimizing uncertainty
- *Control*... accounting for important sources of variation

We will discuss each of these issues in more detail along with other important study design issues in the second part of this course. For now, let's assume the simplest case in which our observational units are scaled perfectly to match our ecological question, we have designed a simple random sampling scheme in which observations are drawn at random from the population to guarantee unbiased parameter estimates, we have ensured a large sample size to minimize uncertainty in our parameter estimates, and we have measured all important sources of variation (i.e., independent causes of variation in the dependent variable) to minimize the unexplained variation in the model. Now, go out and collect the data!